BLAST

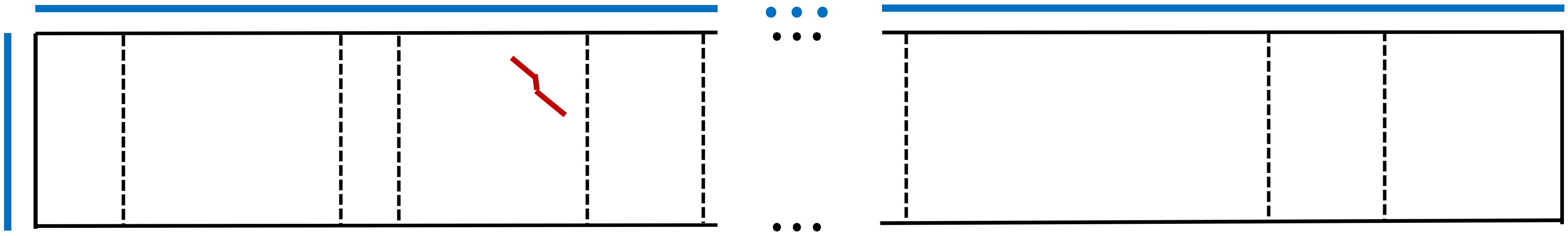
Stephen Altschul

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health

The Problem

The Smith-Waterman algorithm is slower than desirable for database searches.

The SW algorithm spends a lot of time on sequences, and regions of the path graph, that contain no biologically relevant alignments.



There are very fast algorithms for finding perfect matches…

But many interesting alignments are very weak. Biologically relevant protein alignments may contain no runs of more than two identical letters.

BLAST (Basic Local Alignment Search Tool)

BLAST is a widely-used heuristic search algorithm that seeks to approximate the results of the Smith-Waterman algorithm. Like SW, it runs in time $O(mn)$, but with a much smaller leading factor. Typically, BLAST can be expected to run one to two orders of magnitude faster than SW.

Central ideas:

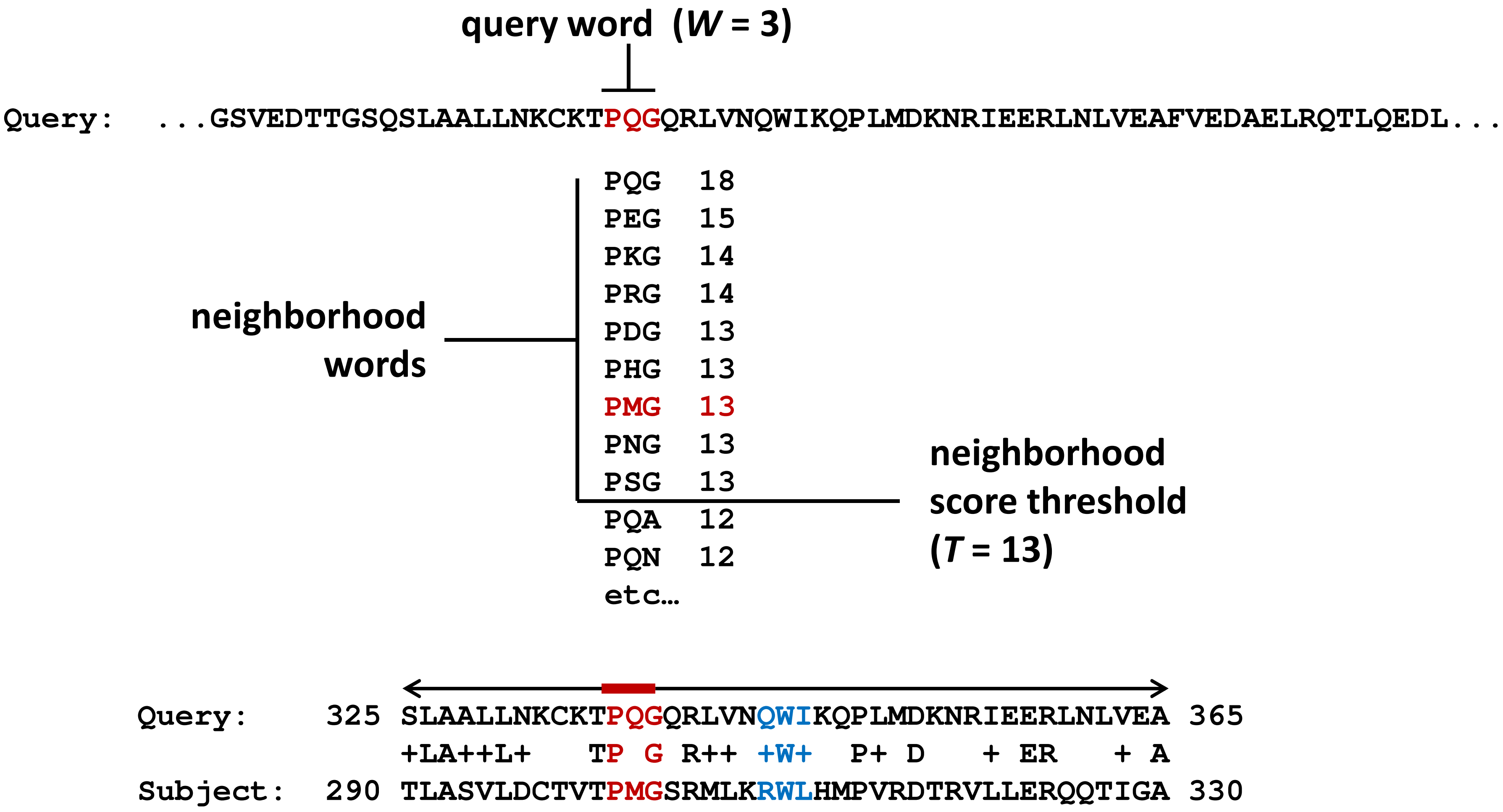
A local alignment with score high enough to be of interest (i.e. to be statistically significant) likely will contain an aligned pair of short *words* with score greater than or equal to some *threshold* value T. We call such a aligned pair of words a *hit*.

To find significant local alignments (called *high-scoring segment pairs*, or *HSP*s), BLAST first seeks hits, and then investigates whether each hit is a chance event, or is contained in an HSP.

For each word in the *query sequence*, BLAST determines what possible words it could pair with to form a hit. BLAST builds these words into a table, and scans the database for exact matches, each of which constitutes a hit.

Altschul, S.F., *et al.* (1990) “Basic local alignment search tool.” *J. Mol. Biol.* **215**:403-410.

The Ungapped BLAST Search Algorithm



query word (*W* = 3)

```
Query:   ...GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL...
```

neighborhood words

PQG	18
PEG	15
PKG	14
PRG	14
PDG	13
PHG	13
PMG	13
PNG	13
PSG	13
PQA	12
PQN	12
etc…	

neighborhood score threshold (*T* = 13)

```
Query:    325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
              +LA++L+   TP G R++ +W+  P+ D   + ER   + A
Subject:  290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
```

High-scoring Segment Pair (HSP)

Finding Hits

As one steps through database sequences, each word may correspond to one or more hits involving words from the query. There are various ways to locate these hits.

Array of lists. Each word can serve as a pointer into an array of lists. The array has size L^W, where L is the size of the alphabet. If many fewer than L^W neighborhood words are generated by the query, the size of the array may be reduced by hashing.

Finite state machine. One may construct a finite state machine to recognize hits. This is generally faster than the list method, but is more complicated.

Mealy, G.H. (1955) "A method for synthesizing sequential circuits." *Bell System Tech. J.* **34**:1045-1079.

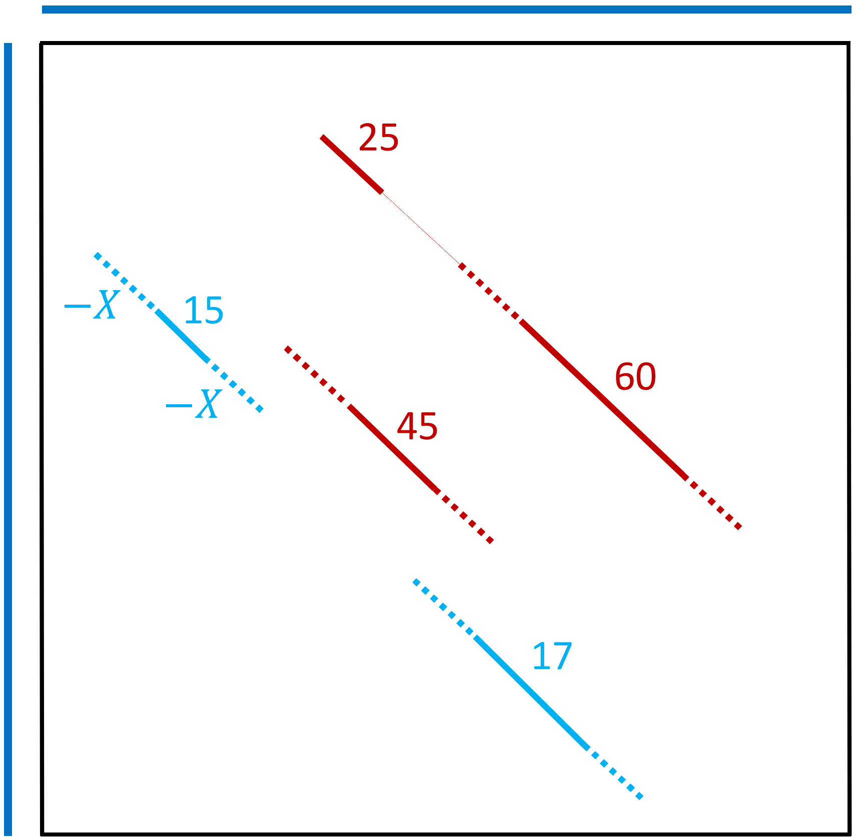
The choice of word length W.

As W increases, so does BLAST's speed for a given level of accuracy, because fewer chance hits need to be examined. However, BLAST's space requirements grow exponentially with W. For protein-protein BLAST, W was originally set to 4, but then reduced to 3 to run on the limited-storage PCs of the day.

The "*X*-drop" Algorithm and Parameter

To be confident of finding the optimal local alignment starting from a hit, one must extend the alignment in both directions to the path graph boundaries. However, usually this will cost a good deal of time for no benefit.

BLAST extends an alignment in each direction until the running score drops *X* below the best score seen so far. Sometimes this can result in a non-optimal alignment being reported.



Tradeoff Between the Threshold Parameter and Speed

The number of words generated by a query sequence decreases as the threshold parameter increases.

The time required by BLAST is approximately a linear function of the number of words it generates.

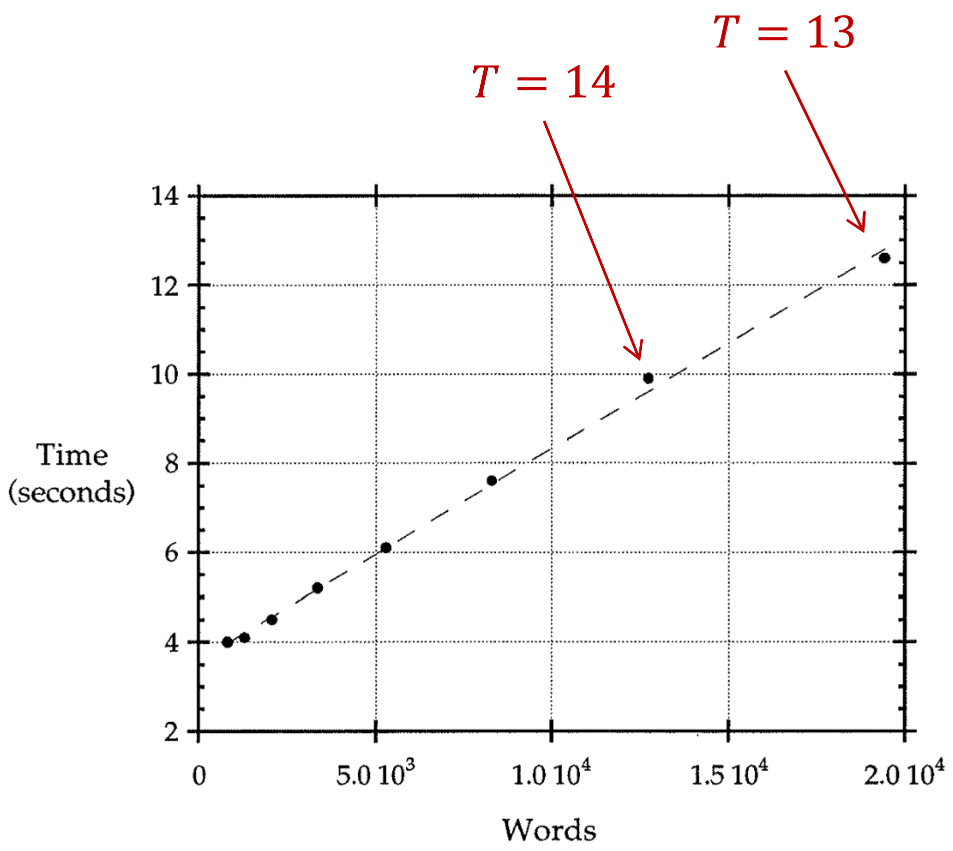


Figure from: Altschul, S.F. *et al.* (1990) *J. Mol. Biol.* **215**:403-410.

Tradeoff Between the Threshold Parameter and Accuracy

Increasing T increases the probability that a weak alignment will be missed.

However, for a fixed T, the probability q of BLAST missing an HSP with score S, decreases exponentially with S, so very strong alignments are likely to be found ever with high values of T.

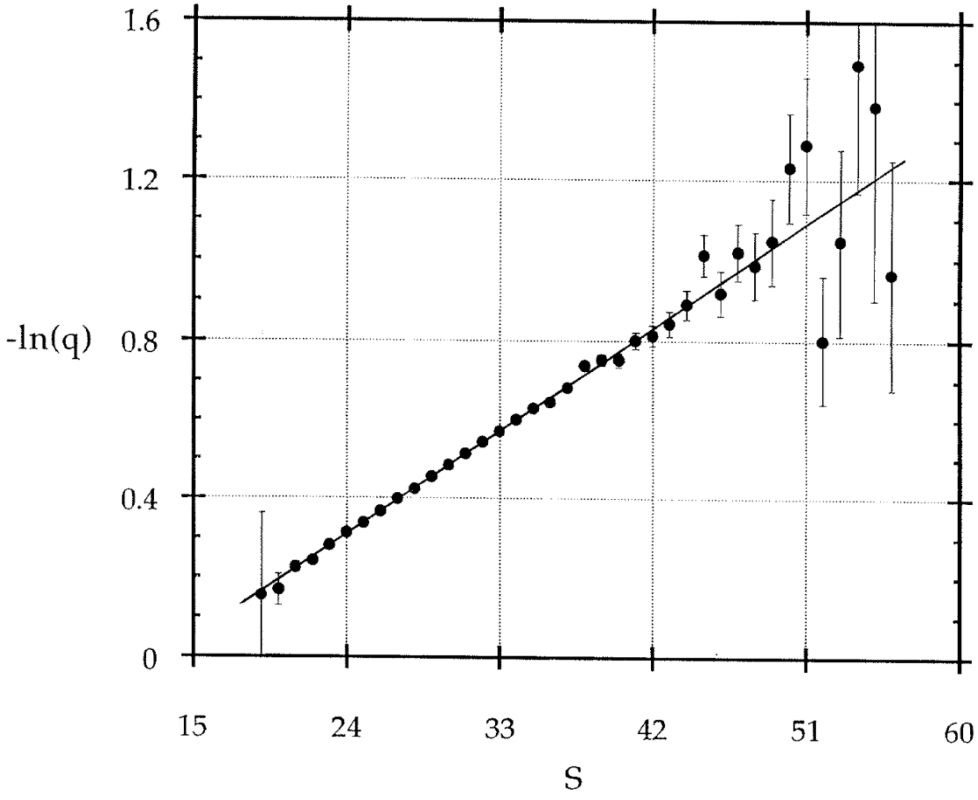


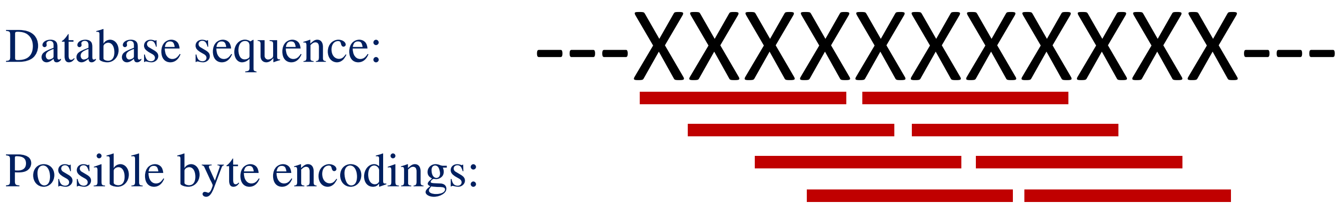
Figure from: Altschul, S.F. *et al*. (1990) *J. Mol. Biol.* **215**:403-410.

BLAST for DNA Sequences

The general method used for proteins is applicable as well to DNA sequences. However, most DNA alignments of interest have fairly long runs of matches. Thus, for DNA database searches, BLAST requires a hit be a run of W matching letters.

Typically, we construct a table with $4^8 = 65{,}536$ pointers (2 bytes), to lists of all occurrences of each 8-word in the query. (Most lists are usually empty.) To find hits, the database is first scanned for 8-words that match a query 8-word, and these are checked to see whether they form part of a run of W matching letters. Any hit found is then extended using the X-drop algorithm.

Setting $W = 11$ has the advantage that, when database DNA sequences are encoded as bytes, any hit implies at least two consecutive matching bytes. Thus one need not "unpack" the database sequence into individual bases until after an 8-word match is found.



Database sequence:

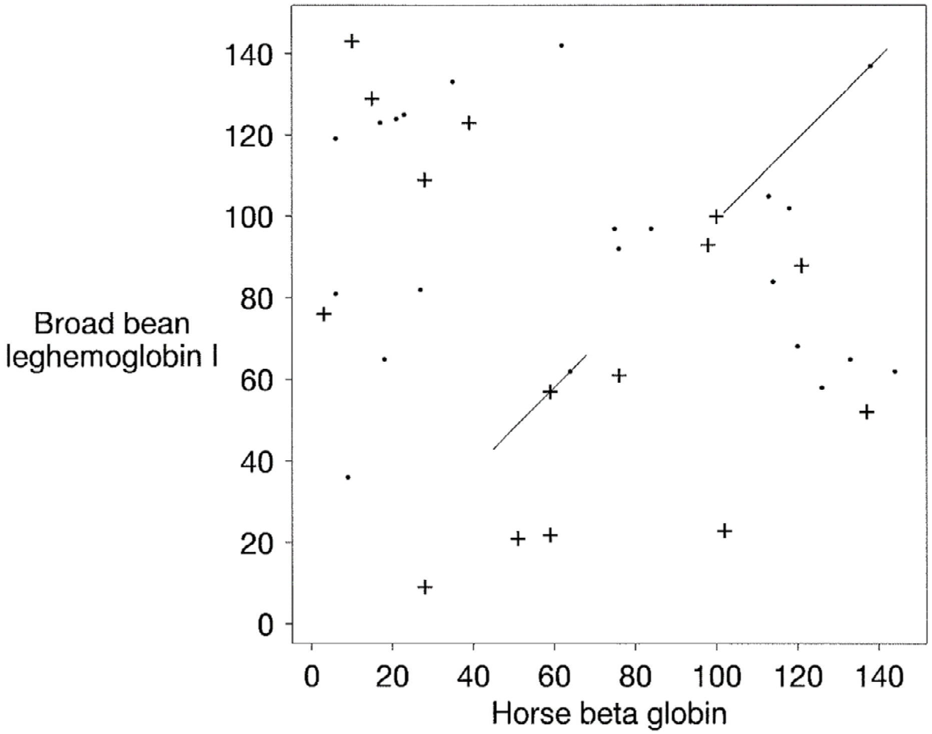
---XXXXXXXXXXX---

Possible byte encodings:

Compensating for No Gaps: Multiple HSPs

For many years, the original ungapped BLAST program partially compensated for its inability to find gapped local alignments by reporting *multiple* ungapped alignments for a pair of sequences, and providing a combined statistical assessment.

Because second-best or third-best HSPs may have relatively low scores, and be harder to find, this requires lowering BLAST's T parameter, and slowing down the program. Here, to find the second best HSP, which increases statistical significance, requires using $T = 11$.



Hits: $T \geq 13$ +
$T \geq 11$ ●

Karlin, S. & Altschul, S.F. (1993) "Applications and statistics for multiple high-scoring segments in molecular sequences." *Proc. Natl. Acad. Sci. USA* **90**:5873-5877.

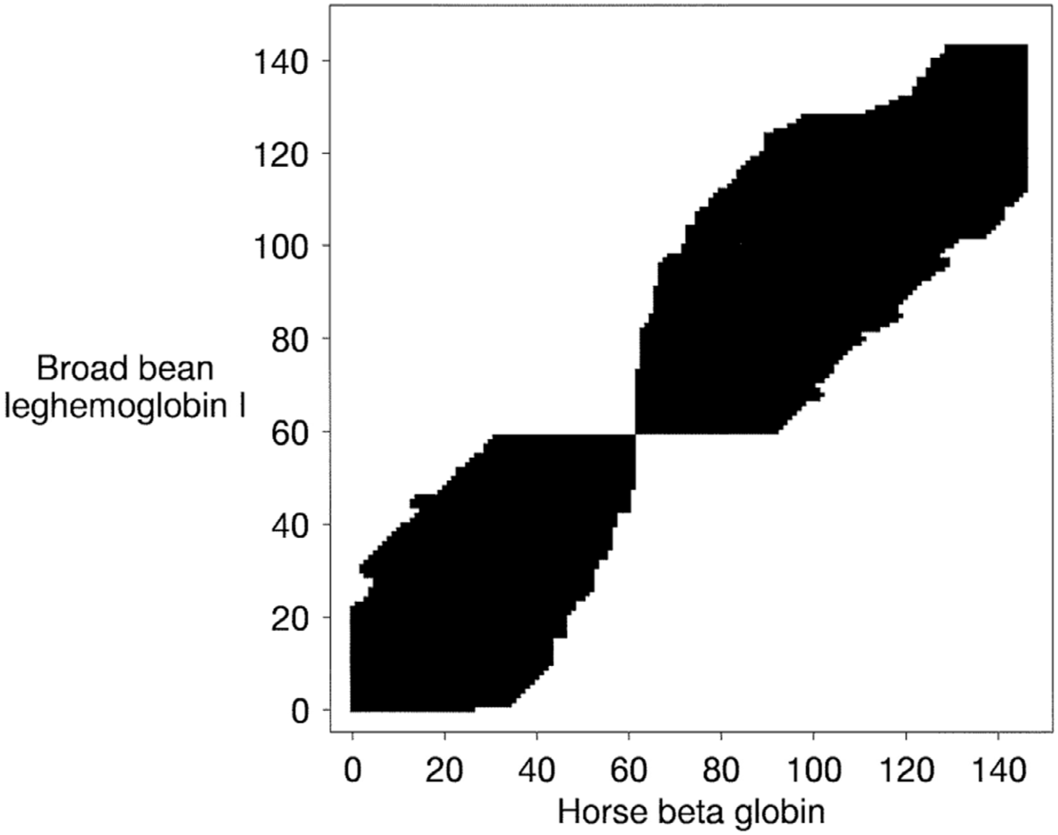
Gapped BLAST

Outline of the gapped BLAST strategy:

1) Invoke a gapped extension whenever an ungapped HSP's score reaches a threshold S_g. Because gapped extensions are very costly, choose S_g so that only about 2% of unrelated database sequences trigger a gapped extension.

2) Seed the gapped extension from the highest-scoring region of the triggering HSP.

3) Generalize the *X*-drop procedure to the Smith-Waterman algorithm.



Altschul, S.F. *et al.* (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Res.* **25**:3389-3402.

Gapped BLAST continued

Note that most gapped extensions are still triggered by chance HSPs that are not part of a higher-scoring gapped alignment. These gapped extensions will tend to be of limited extent.

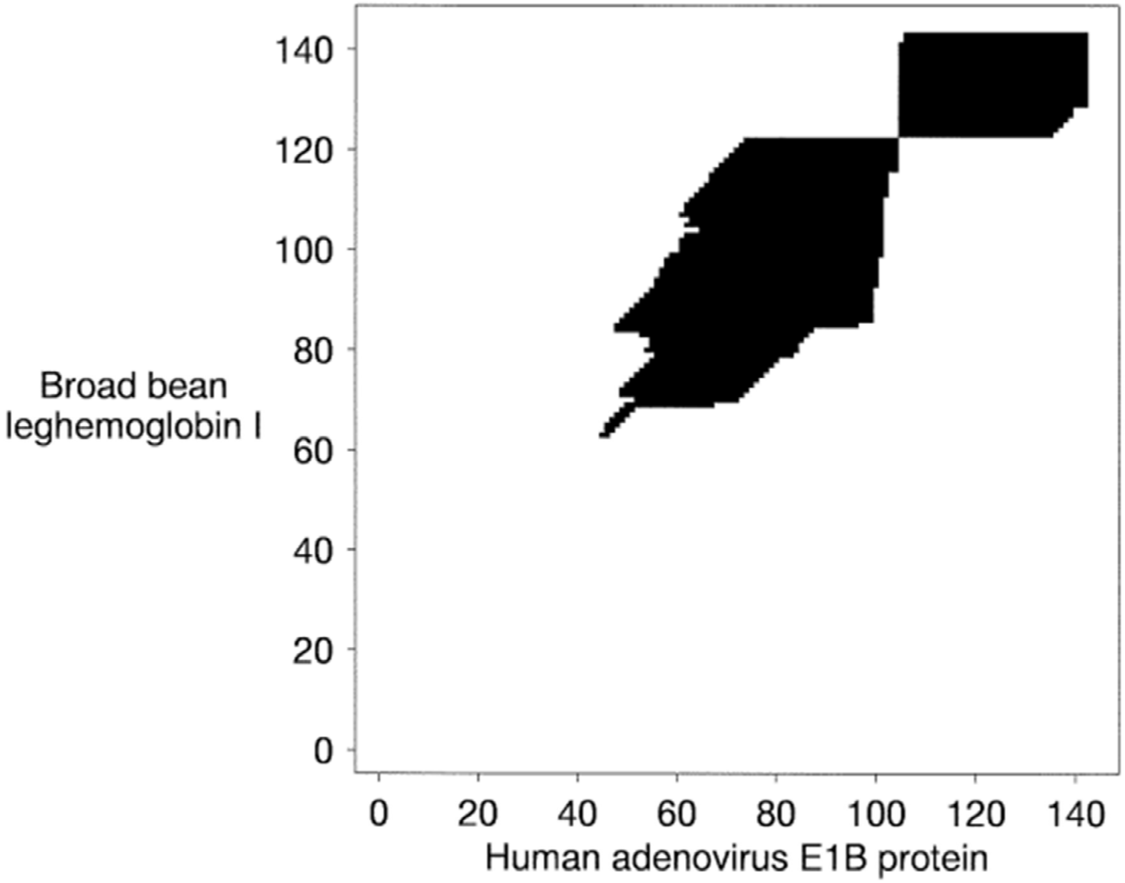
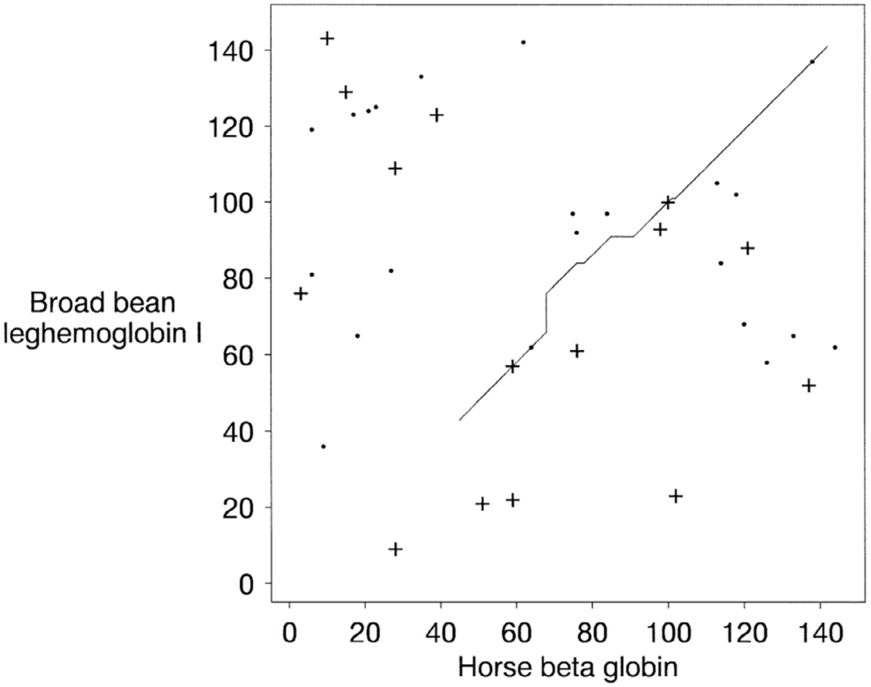


Figure from: Altschul, S.F. *et al.* (1997) *Nucleic Acids Res.* **25**:3389-3402.

Gapped BLAST continued

The path of the optimal local alignment:



```
Leghemoglobin  43 FSFLKDSAGVVDSPKLGAHAEKVFGMVRDSAVQLRATGEVV--LDGKDGS------ 90
                  F  L +   V+ +PK+ AH +KV          L + GE V  LD   G+
Beta globin    45 FGDLSNPGAVMGNPKVKAHGKKV----------LHSFGEGVHHLDNLKGTFAALSE 90

Leghemoglogin  91 IHIQKGVLDP-HFVVVKEALLKTIKEASGDKWSEELSAAWEVAYDGLATAI 140
                  +H  K  +DP +F ++   L+  +    G  ++ EL A+++    G+A A+
Beta globin    91 LHCDKLHVDPENFRLLGNVLVVVLARHFGKGFTPELQASYQKVVAGVANAL 141
```

Figures from: Altschul, S.F. *et al.* (1997) *Nucleic Acids Res.* **25**:3389-3402.

Gapped BLAST Speedup: The Two-Hit Method

Two hit strategy: Lower the threshold T to allow more hits, but extend only if two hits fall on the same diagonal, and within a window of fixed length. In this example, lowering the threshold from $T = 13$ (+) to $T = 11$(•) increases the number of hits from 15 to 37, but decreases the number of extensions from 15 to 2.

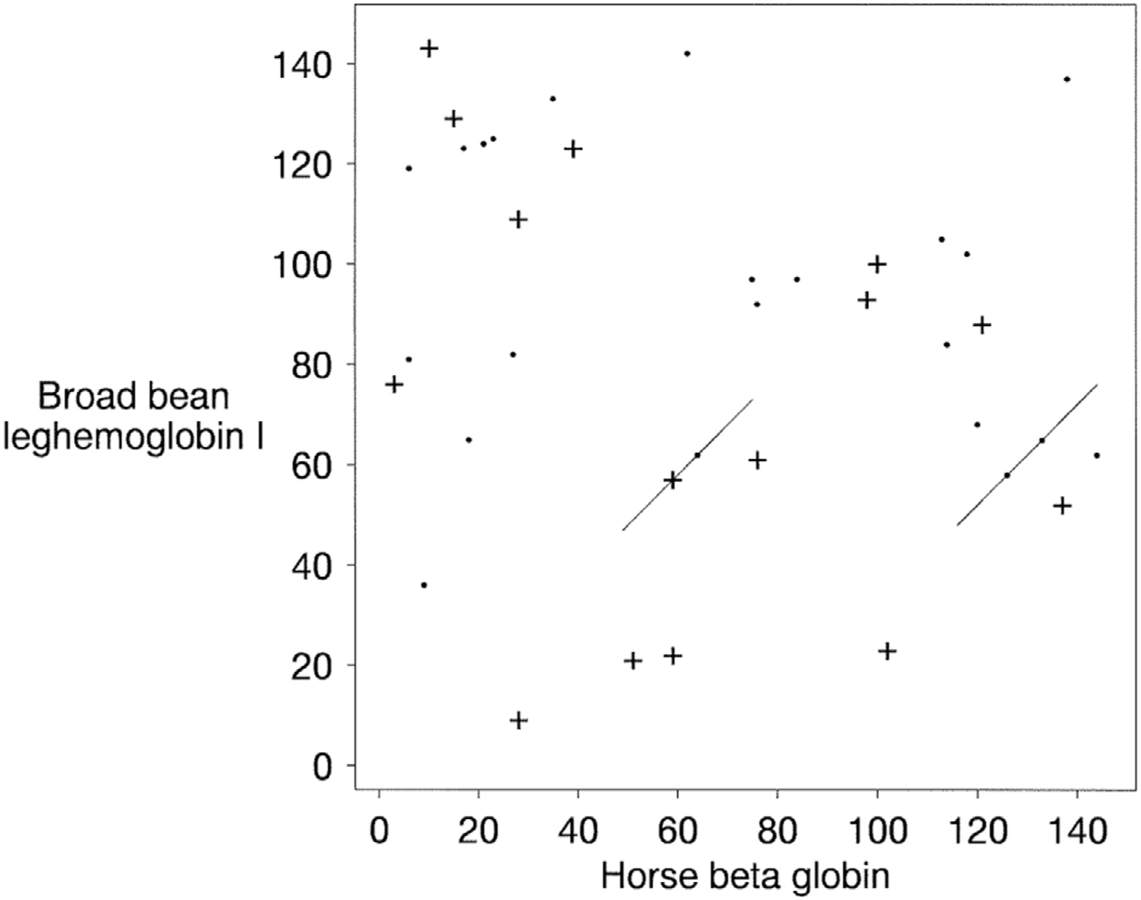


Figure from: Altschul, S.F. *et al.* (1997) *Nucleic Acids Res.* **25**:3389-3402.

The Two-Hit Method: Accuracy

The two-hit method increases program speed, because the gain from the smaller number of "costly" extensions outweighs the loss from the increased number of "cheap" hits.

The program's accuracy improves for high-scoring HSPs, but is worse for low-scoring HSPs. Thus this method is not appropriate for finding multiple HSPs, but it is appropriate for gapped BLAST.

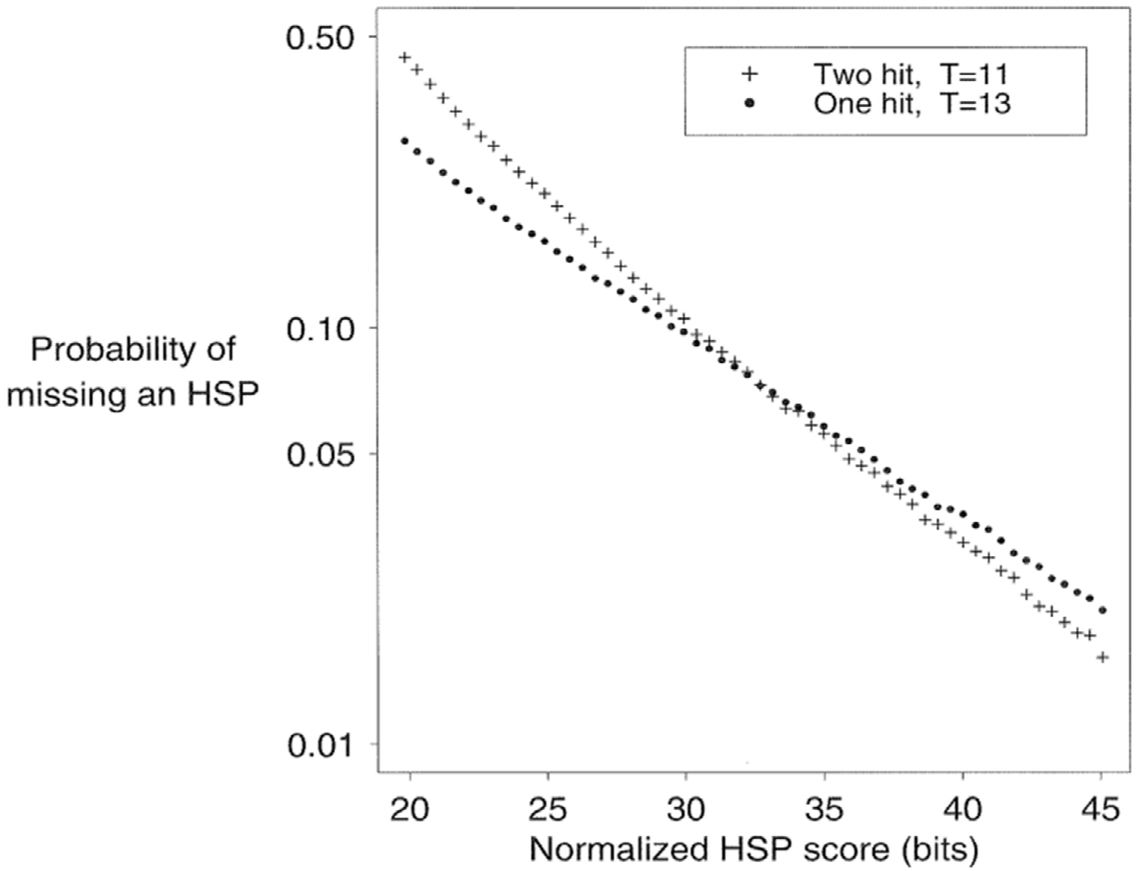


Figure from: Altschul, S.F. *et al.* (1997) *Nucleic Acids Res.* **25**:3389-3402.

Relative times spent by the original and gapped BLAST programs on various algorithmic stages

	Overhead: database scanning, output, etc.	Calculating whether hits qualify for ungapped extension	Ungapped extensions	Gapped extensions	Total
Original BLAST:	8 (8%)		92 (92%)		100 (100%)
Gapped BLAST:	8 (24%)	12 (37%)	5 (15%)	8 (24%)	33 (100%)

Table from: Altschul, S.F. *et al.* (1997) *Nucleic Acids Res.* **25**:3389-3402.

Varieties of BLAST

The BLAST programs are available on the NCBI web site:
http://blast.ncbi.nlm.nih.gov/

Different applications call for different versions of the program:

blastn: Search a nucleotide database using a nucleotide query

blastp: Search a protein database using a protein query

blastx: Search a protein database using a translated nucleotide query

Used to look for exons in cDNA sequences.

tblastn: Search a translated nucleotide database using a protein query

tblastx: Search a translated nucleotide database using a translated nucleotide query

Infrequently used.

Some Other Heuristic Local Alignment Algorithms

FASTA

FASTA was the most widely used DNA and protein sequence database search program before the advent of BLAST. It is similar to BLAST in many ways, and is still frequently used. Like BLAST, it is a heuristic for approximating the Smith-Waterman algorithm, but uses different heuristic methods to increase speed. BLAST and FASTA also use somewhat different methods to calculate statistical significance.

Pearson, W.R. & Lipman, D.J. (1988) "Improved tools for biological sequence comparison." *Proc. Natl. Acad. Sci. USA* **85**:2444-2448.

Indexed BLAST

When a large DNA database, such as a completed genome, is fixed, and will be searched many times, one can index the database, and devote a machine to keeping the index in memory. A query sequence can then simply look up hits in this index. This basically reverses the roles of the query and database.

BLAT

Kent, W.J. (2002) "BLAT--The BLAST-Like Alignment Tool." *Genome Res.* **12**: 656-664.